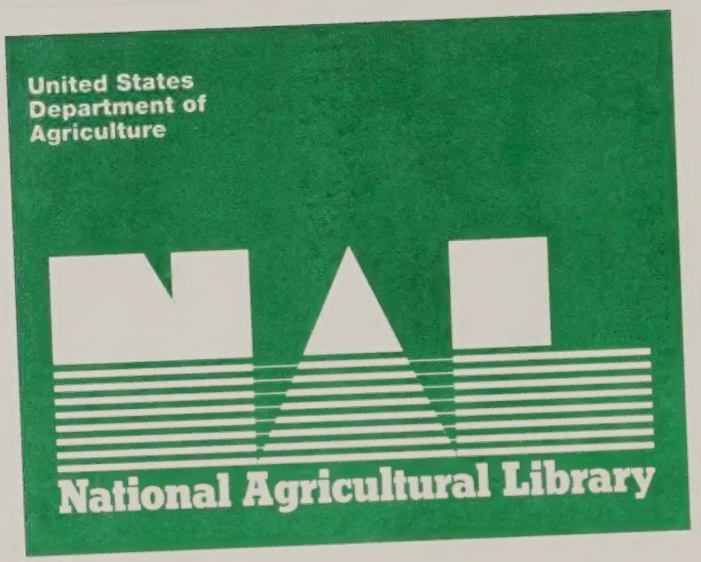
Data Resources and the Plant Genome Research Program

A report by the
Plant Genome Database Subcommittee
Science and Technology Coordinating Committee
USDA Plant Genome Research Program

United States Department of Agriculture
Beltsville, Maryland
June 1990



For additional copies of this report contact: The National Agricultural Library, Room 100, 10301 Baltimore Boulevard, Beltsville, Maryland 20705. (301) 344-3834

TABLE OF CONTENTS

INTRODUCTION

This report supports planning for the USDA Plant Genome Research Program, a national research program announced by Secretary of Agriculture Clayton Yeutter in February 1989. Since that announcement, (1) a program director has been named; (2) a Science and Technology Coordinating Committee, co-chaired by representatives from the Agricultural Research Service and the Cooperative State Research Service, has been organized to help develop the program; and (3) a subcommittee of that Committee (see Appendix A) was appointed to review related database efforts, identify other sources of information, and assemble information on user needs related to the genome research program. The work of that subcommittee to date is reported herein.

This document is a specialized resource that pulls together genome related information from a variety of sources. It builds upon earlier and ongoing work of (1) the original conferees who attended a December 1988 meeting to begin planning this effort, (2) the Science and Technology Coordinating Committee, and (3) other organizations and individuals. It lays the foundation for future plant genome data organization and management. It focuses on the current status of data management in genome mapping efforts, and on where such efforts seem to be headed. It also begins the process of identifying the types of genome related information needed by plant breeders and other researchers.

References

Plant Genome Research Conference Report. Proceedings of the Crop and Forest Genome Mapping Conference, Washington, D.C., December 12-14, 1988. Washington, D.C.: United States Department of Agriculture, Science and Education.

USDA Plant Genome Mapping Program, Science and Technology Coordinating Committee: August 30-31, 1989 Meeting Summary. Washington, D.C.: United States Department of Agriculture, Agricultural Research Service. December 1989.

CURRENT VIEWS ON GENOME INFORMATICS

Data analysis and management are among the major concerns of both the Plant Genome Database Subcommittee and the Human Genome Joint Informatics Task Force (JITF). Many of the concerns identified by JITF (or its predecessor committees that advised the National Institutes of Health and the Department of Energy on human genome data-related issues) are relevant to plant genome informatics.

Applications and needs which are essentially the same for both plant and human genomes include:

1. storage, retrieval, and analysis of nucleic acid and protein sequences;
2. construction of high and low resolution physical maps;
3. development of electronic laboratory notebooks to handle large amounts of experimental data;
4. ability to gain access to information on a variety of platforms by persons with different levels of computer expertise;
5. support for automated sequencing and other large scale sequencing and mapping efforts;
6. development of links between various databases and tools for comparison of data from different sources; and
7. standardization of data exchange and communication network protocols and integration of database schemata.

Existing national and international data resources and software currently in use by investigators of all genome projects include:

1. nucleotide sequence databases, especially GenBank, the European Molecular Biology Laboratory (EMBL) Data Library, and the DNA Data Bank of Japan (DDBJ);
2. protein sequence databases such as the National Biomedical Research Foundation-Protein Identification Resource Protein Sequence Database (NBRF-PIR) and the SWISS-PROT Protein Sequence Data Bank (SWISS-PROT);
3. protein structural databases, e.g., Protein Data Bank (PDB) at Brookhaven National Laboratories, and the Cambridge Structural Database (CSD);
4. bibliographic databases such as Medline, Biosis Previews, AGRICOLA, and the electronic version of Chemical Abstracts;

5. software tools from commercial vendors such as IntelliGenetics, Genetics Computer Group (formerly UWGCG, now a private company), International Biotechnologies, Inc., and DNASTAR;

6. NIH supported research resources such as the Molecular Biology Computer Research Resource at the Dana-Farber Cancer Institute, Boston, which distributes public domain analysis software for the analysis of DNA, RNA and protein sequences; and

7. the Listing of Molecular Biology Databases (LiMB) and the Directory of Biotechnology Information/Resources (DBIR), two electronic directories to sources in biotechnology and molecular biology.

In addition to these existing sources, new resources likely to be widely applicable include electronic notebooks under development at the Los Alamos National Laboratory and the Lawrence Berkeley Laboratory, and the database system under development at the National Library of Medicine's National Center for Biotechnology Information.

Although general data analysis and management needs are the same for both human and plant genome informatics, there are several areas in which plant genome research and informatics differ significantly from the Human Genome Project.

1. The problem of database design incompatibilities resulting from alternate representation of similar concepts has been recognized by JITF. This problem is more acute for the diverse databases contemplated for the plant genome.

2. There exists much genotype information about many genetic strains of a wide variety of organisms. Availability of material from germplasm banks and genetic stock centers is a vital component of this information. Support for facilities, datasets and the development and augmentation of computerized databases is essential to plant genome informatics.

3. The plant genome databases must support genetic descriptions of diverse organisms with different genetic systems and experimental histories. A plant species may be monoecious or dioecious, apomictic, self-fertile or outcrossing, amenable to single cell cloning and regeneration, haploid, diploid, triploid, or polyploid in some or all tissues, an angiosperm or a gymnosperm, with life-cycle of weeks or centuries. The plant genome databases also must accomodate a variety of research approaches to different genetic questions. Plant research provides information in the fields of cytogenetics; molecular biology; Mendelian genetics; practical breeding; and physiological, population, ecological or evolutionary genetics. Representing this information, assessing user needs for access to such information, and developing appropriate access mechanisms are major challenges for biological information management.

4. The emphasis on cross species comparisons of genes affecting agriculturally important traits such as drought resistance, disease resistance, symbiotic interactions, or storage protein or carbohydrate biosynthesis requires simultaneous access to data from a multitude of databases.

5. The plant research community includes private seed, biotechnology, and food processing companies; laboratories at research universities; extension programs; commodity group associations; stock centers and germplasm repositories; and international research centers. The diversity of this community, which is diffused throughout the public and private sectors, complicates efforts to identify resources, coordinate research, and develop collaborative databases.

Mechanisms such as interdisciplinary degrees and senior fellowship programs are needed to promote collaborative project funding and enhance communication between computer scientists and plant biologists. They would aid the development of a pool of scientists with appreciation of both areas as well as expertise in at least one. These scientists would then have the interdisciplinary expertise required to develop needed systems. It is unlikely that training programs fostered by the Human Genome project will adequately encompass agricultural and plant biology, therefore support from the Plant Genome Research Program is recommended.

A pilot program involving two or three plant species would be very useful to identify the existing databases and develop the software to link information from all databases about two or more species. From this project, problems of database structure and management could be predicted.

CURRENT STATUS OF DATA RESOURCES RELATED TO THE PLANT GENOME RESEARCH PROGRAM

A few of the primary existing databases for protein or DNA sequences and for molecular biology information sources are summarized in this section. More details are provided in Appendix C. Trends in development of sequence databases are identified because consideration of these may provide insight useful to the design of new databases and delivery systems. Databases for RFLP datasets and maps are less well developed. Trends in the development of these are identified in this section. Appendix B identifies some of the diverse datasets and databases available to corn researchers. These resources are more developed for corn than for other crops.

Nucleotide Sequence Databases

A single database has emerged as the primary repository of nucleotide sequence data generated in the United States and, through its collaborations, data generated worldwide. The GenBank Nucleotide Sequence Data Bank (GenBank) has as its mission the collection, organization, maintenance, and distribution of all published nucleotide sequences. Additional sequence information in the database includes bibliographic citations, experimental source of the sequenced genetic material, taxonomic classification of the source organism, mapping of the sequence on the organism's genome, and the function of the sequence. Inclusion of these data and the format of the database permit a variety of approaches to searching the database in addition to sequence matching (Foley *et al.*, 1986). In March 1990 (Release 63.0), GenBank contained 33,377 sequence entries representing 40,127,752 bases. Of these, 2,332 sequence entries containing 3,501,265 bases originated from plant sources (Moore *et al.*, 1989).

GenBank is distributed in quarterly releases on magnetic tape and semiannual releases on floppy diskettes. It is available online with daily updates. Releases on CD-ROM are planned for Spring 1990. GenBank is funded by the U.S. National Institute of General Medical Sciences (NIGMS) and is currently co-sponsored by the National Library of Medicine and the U.S. Department of Energy.

During 1988 alone, GenBank grew by 47%. Growth of the database in recent years has been nearly exponential with a doubling time of about 16 months (Moore *et al.*, 1989). There are several mechanisms through which GenBank attempts to achieve the comprehensiveness with which it is charged:

1. GenBank works closely with the EMBL Data Library (EMBL) and the DNA Data Bank of Japan (DDBJ), its European and Asian counterparts. Through this

collaboration, data from the three databanks are shared. GenBank and EMBL currently share releases. Data from DDBJ are added to GenBank releases as available.

2. GenBank, EMBL, and DDBJ regularly scan the major journals that publish sequence data to identify sequences for the database.
3. A means of direct submission of entries by authors has been implemented. New software to facilitate this entry, *Authorin*, has been developed and is being distributed.
4. Some journals now submit sequence data to GenBank or require evidence that sequences have been directly submitted by authors.
5. A joint "missing data" project among GenBank, EMBL, and DDBJ has identified sequences published before October 1, 1988 that were not previously included in the data banks.

Protein sequence databases

The major protein sequence database in the United States is the National Biomedical Research Foundation-Protein Identification Resource Sequence Database (NBRF-PIR Protein Sequence Database). Production of the database by the National Biomedical Research Foundation (NBRF) is currently supported by a grant from the Extramural Division, National Library of Medicine. This database is one component of the Protein Identification Resource (NBRF-PIR), which includes access to protein and nucleotide sequence databases and software tools developed to facilitate access and analysis. The NBRF-PIR Protein Sequence Database itself includes both sequenced proteins and translated nucleic acid sequences as well as associated data such as bibliographic citations, annotations indicating post-translational modifications, etc., and availability of X-ray crystallography and/or nucleic acid sequence data (Sidman *et al.*, 1988). The contents of the database are carefully reviewed, edited, and annotated by scientists at NBRF (Barker *et al.*, 1987).

Collaborative efforts similar to those among nucleotide sequence databases have developed among the major protein sequence databases. PIR-International is the official coordinating body for NBRF-PIR Protein Sequence Database (USA), JIPIDS (Japan) and MIPS (FRG) (Tsugita, 1989). Each serves as a collection node for sequence data in its geographic area, but the records generated in this way are shared among the databases.

The SWISS-PROT database is a protein sequence collection whose format is compatible with the EMBL Nucleotide Sequence Data Library. The database is built by reformatting data from the NBRF-PIR Protein Sequence Database and by translating the EMBL Nucleotide Sequence Database. Developed by Amos Bairoch of the University of Geneva, SWISS-PROT is now produced and distributed by EMBL.

Trends in the development of sequence databases

Several important trends have become apparent in the development of sequence databases:

1. A single large database may serve a geographic area. Collaboration among databases in different geographic areas and between databases for nucleotide sequences and protein sequences within an area has increased. Collaborative efforts have resulted in a common entry format, sharing of journals scanned, and increasingly rapid data exchange between databases.
2. Use of print formats has been phased out. Distribution using floppy diskettes is becoming unwieldly as the size of the databases increases. There is a trend toward higher capacity formats such as CD-ROM. Online access and mail servers are becoming available to address the problem of current data.
3. The time lag between sequence determination in the laboratory and reporting of the data in the database has been reduced by a variety of techniques including direct entry of data by authors, and cooperation of journals in submission of entries.
4. Database developers are incorporating "pointers" within records that inform users of other databases containing related records.
5. As print formats (flat files) are phased out for data maintenance and distribution, the sequence databases are being restructured as relational databases. Among the advantages of this change are the ability to use commercial database management software and increased data integrity which results from structuring the data according to a theoretically well-understood data model.

Restriction Fragment Length Polymorphism (RFLP) datasets and maps

Restriction fragment length polymorphisms (RFLPs), first described in 1980, now play a major role in human genetic research and the mapping of the human genome. Over the past few years, RFLP technology has been applied to important crop species. RFLP maps are now available for at least seven different crop species (Tanksley *et al.*, 1989). Appendix D identifies a few of the major public sector research programs by crop.

The potential use of RFLPs to accelerate plant breeding programs has been recognized by industry. For example, Agrigenetics has produced a map of the corn genome with over 300 markers (Ratner, 1989). NPI (Salt Lake City, Utah) has developed maps of corn, tomato, onion, *Brassica oleracea*, *B. campestris*, *B. napus* (partial map), and is developing maps for other crops (Allred, NPI, personal communication). Five European seed companies are funding the development of RFLP maps for barley and wheat at the Agricultural and Food Research Council (AFRC) Institute for Plant Science Research (IPSR) in Cambridge, U.K. (Newmark, 1989). A preliminary list of some companies known to have RFLP capabilities is included in Appendix D. RFLP

probes, datasets, and maps generated in this manner may be proprietary and generally unavailable or available at high cost. For example, access to the RFLP probes and maps generated at IPSR will be restricted for three years to the funding seed companies (Newmark, 1989).

In contrast to nucleotide and protein sequence data, data being generated by RFLP activities are scattered, may not be accessible to the public, and are being recorded in a variety of databases being generated by the researchers. A list indicating availability should be constructed and made available to researchers.

Sources of information about databases pertinent to the plant genome program

The Directory of Biotechnology Information/Resources (DBIR) at the National Library of Medicine serves as an online directory to publicly available biotechnology information, including databases. The Listing of Molecular Biology Databases (LiMB) identifies and characterizes databases relevant to molecular biology. These sources are described in more detail in Appendix C.

The National Plant Germplasm Committee issued a special report which gives an overview of the National Plant Germplasm System. The report includes recommendations related to genetic stocks collections. Brief descriptions of existing genetic stock collections are included by commodity. See Appendix E.

Appendix F is a list of genetic stock collections for crop species. The list was provided by the Office of Germplasm Resources Information Network/USDA.

Other sources of information pertinent to the plant genome program

The Center for High Performance Computing, University of Texas System, is compiling a bibliography in the field of computational genetics focusing on computer and mathematical aspects of genetics. A first draft of the bibliography, containing 611 citations, was released in April 1990. Sarah Barron, one of the authors of that bibliography, reports that a group in Switzerland has been compiling a similar, but more extensive bibliography, since 1987. The Sequence Analysis Bibliographic Reference Databank is available through the European Molecular Biology Laboratory (EMBL) database. These are valuable resources to the Plant Genome Research Program.

Germplasm Resources Information Network

In order to promote the continued improvement of agricultural crops, the U.S. National Plant Germplasm System (NPGS) has as its mission the collection, documentation, preservation, evaluation, enhancement, and distribution of plant genetic resources. It is critical that management of these genetic resources include the management of information related to the plant materials being preserved. The Germplasm Resources Information Network (GRIN) is the information management component of the

NPGS. GRIN is a centralized computer database that serves as a repository of information about the plant genetic resources within the NPGS and provides access to this information for users of the system. The database is more fully described in Appendix G.

References

Foley, Brian T., Debra Nelson, Maura T. Smith, and Christian Burks. 1986. Cross-sections of the GenBank database. *Trends in Genetics* 2 (9): 233-238.

Moore, John F., David Benton, and Christian Burks. 1989. The GenBank Nucleic Acid Database. *Focus* 11 (4): 69-72.

Sidman, Kathryn E., David G. George, Winona C. Barker, and Lois T. Hunt. 1988. The Protein Identification Resource (PIR). *Nucleic Acids Research* 16 (5, Part A): 1869-1871.

Barker, W. C., D. G. George, and M. C. Blomquist. 1987. The Protein Identification Resource (PIR). In *Biotechnology Information '86:Proceedings of a conference held at Sussex University, United Kingdom, 22-25 September 1986*, edited by Richard Wakeford, 35-47. Oxford, England: IRL Pewaa Limited.

Tsugita, Akira. 1989. Current Status of Protein Data Banks. In *Methods of Protein Sequence Analysis: Proceedings of the 7th International Conference, Berlin, July 3-8, 1988*, edited by Brigitte Wittmann-Liebold, 361-364. New York: Springer-Verlag.

Tanksley, S. D., N. D. Young, A. H. Paterson, and M. W. Bonierbale. 1989. RFLP Mapping in Plant Breeding: New Tools For an Old Science. *Bio/Technology* 7: 257-264.

Newmark, Peter. 1989. Serial RFLPs For Cereal Grains. *Bio/Technology* 7: 211.

Ratner, Mark. 1989. Crop Biotech '89: Research Efforts are Market Driven. *Bio/Technology* 7(3): 337-341.

DATABASE NEEDS OF USERS OF PLANT GENOME DATA

Members of the Plant Genome Database Subcommittee informally discussed database needs and potential uses with scientists who are involved in plant genome research in both the public and private sectors. These interviews do not represent an official survey or wide sampling, and the Subcommittee strongly recommends that a more complete assessment of users' needs be made at the outset and monitored throughout the course of database development. However, until more comprehensive information is available, these comments can alert planners in the USDA Plant Genome Research Program to some of the perceived needs and concerns presently felt by the plant genome research community.

Users' needs are addressed at four levels: the general problems that plant breeders, basic researchers (in genetics, molecular biology, physiology, other fields), and others want to solve, using information resident in current or future plant genome databases; cloning queries; mapping and genetic characterization queries; and germplasm/ pedigree queries. A fifth section includes a few general comments about the databases required.

I. Types of problems presented (categorized according to type of genomic information required)

A. Problems using specific RFLP markers to track desirable or undesirable traits.

1. Tracking heterozygous progeny. Breeders can use an RFLP marker to detect heterozygotes carrying a desirable recessive trait without having to do the time consuming backcrosses after each selfing or other cross.

2. Culling seedlots. An RFLP marker associated with an undesirable trait such as viral susceptibility can be used to screen and discard seedlots with high frequency of susceptibility as an alternative to extensive breeding to eliminate the trait.

3. Selecting for desirable traits. An RFLP marker associated with a desirable trait can be used to screen and select seedlots with high frequencies as an alternative to extensive selective breeding.

B. Problems using general RFLP patterns to assess ancestry.

1. Determining parentage of a desirable plant with only partially characterized ancestry.

a. to reproduce or mimic the crosses leading to the desirable line.

b. to survey genetic relatedness of uncharacterized plants in a population containing a known commercially desirable plant. This can be useful, for example, for developing non-native plants imported prior to changes in export laws that no longer allow introduction of a species or current commercial variety.

2. Determining whether two lines from different sources are in fact genetically identical. This may be used to resolve legal disputes between the different sources involving independent origin versus illegal usage.

3. Homozygosity (completeness-of-backcross) assessments. After backcrossing hybrid AB with B for x generations, how much A background is left in the individuals?

4. Determining probable origin of successful plants in a naturally growing or selected population, e.g., planted or wild invaders.

5. Eliminating contaminated hybrid seedlots. Determination of RFLP pattern can identify hybrid seedlots containing high levels of parental inbred seed.

C. General scientific concerns

1. Reconciling the RFLP and genetic maps of a species at increasingly high levels of resolution, i.e., establishing colinearity between the genetic/morphological/cytological map and RFLP maps. Because of the preservation of plant seed stocks, the opportunity exists for redoing and rescoring crosses underlying the classical data as well as repeating RFLP analyses.

2. Determining synteny between different species: Is there conservation of map order of related genes at the interspecies level for genes related by sequence homology as well as at the physiological level?

3. Facilitating analysis and data access for quantitative traits and analyzing the role of quantitative inheritance in describing important agronomic traits.

4. Developing databases that provide access to information on important agricultural characteristics, including geographic, environmental and controlled experimental information. Cited examples of important agricultural characteristics, which will vary among crops, include yield, ripening/maturity data, percent oil, percent soluble solids, pest

resistance, cold hardiness, drought resistance, etc. Including the statistical parameters characterizing these traits is a desired aspect of such databases.

II. Cloning Queries

A. Laboratory bench-level queries for available sequence information. Sequencing laboratories working on specific genes, pathways, or chromosome regions use standard sequence and homology queries handled by existing software querying against GenBank and PIR sequences. No plant-specific deficiencies in annotation in these databases were cited. The laboratories contacted do not use GRIN or Stock Centers and identified no special needs for new types of databases. Use of these databases is discussed in types of queries presented below. It was suggested that RFLP maps and associated software be more accessible.

B. Queries about specific clones and probes. This class of queries essentially involves tracking what is being done to the clones or other genetic sequence intervals, identifying their location on physical maps and genetic maps, and locating other bibliographic or sequence information about the clone.

1. For a specific cloned gene, what type of clone was used and what was the function of the gene? The latter question links gene function with a genetic map database. The former characterizes the vector, the DNA source, the host, the enzymes used, and the insert size and restriction map.

2. Given the species from which this gene was cloned, for what other species are there clones of similar genes or sequences? For what other species has this clone been utilized to identify homologous genes? Has cloning of any gene in a family related to the species of interest proven useful for detecting homologous genes in this species? For example, other Gramineae clones might be useful to someone beginning to map sorghum. Are there clones in other species of this family that have shown utility for use in sorghum?

3. What other information is available about this clone and its availability? This involves links to bibliographic information, other databases, and persons.

4. Can the restriction map of the cloned DNA be located on a comprehensive physical map and genetic map? This involves links to map databases (p. 7-9).

III. Mapping and Genetic Characterization Queries

A. Where is a specific clone located on a physical map or on a genetic map? What genes or clones are the nearest neighbors to this clone? These questions require the ability to zoom in on the maps with progressively finer detail.

B. Show the cytogenetic features and gene locations of the standard chromosome set or of a specific arm or region of a chromosome. Show the structure of translocated chromosomes involving specific chromosome arms. Where is the breakpoint? Which genes are translocated? Where on the RFLP map is the break located? Similar questions apply to inversions and duplications.

C. Are there viable deletions for a specific gene or specific chromosomal region? Monosomics or trisomics for the chromosome carrying that gene? This introduces a link to the strains and pedigree databases.

D. What is the spectrum of chromosomal rearrangements across populations?

E. What product is encoded by this particular gene? What gene encodes this enzyme or other product?

F. Find all alleles at a given locus and list their phenotypic effects.

G. Show interactions between specific mutations in control loci and affected structural genes.

H. What genes are known in this species that cause a given mutant phenotype? Are there mutants with similar phenotype in other plant species and what is their genotype? This is important to a geneticist postulating a phenotype for a mutation of a specific type or affecting a specific trait in order to screen for the mutation of interest. It is manifested as a direct germplasm query also: what mutant lines have that phenotype?

I. Using a panel of inbred lines linked to germplasm database, do RFLP mapping of a particular mutation.

J. Do the RFLP correlations indicate additional genes that modify the phenotype under study?

IV. Pedigree and Germplasm Queries

A. What mutant lines in a collection manifest a given phenotype and what is the seed source of those lines?

B. What inbred lines are being used for RFLP mapping?

C. Find a strain that is monosomic for a specific chromosome. Provide a phenotypic description of the monosomic strain. The same queries are important for trisomics.

D. Are there polyploid series for this species and closely related species? What are the phenotypic differences?

E. Show the available lines, pedigrees, genotype, and source for specific translocation heterozygotes or monosomics or specific-mutation carriers.

F. For a given cultivar, what is the cultivar type, e.g., isolate, inbred line, hybrid, mutant, etc.? What is the pedigree? What is the source address and contact person where the stock is maintained? See Appendix G for a description of the GRIN database.

V. Some General Remarks About the Perceived Needs For Different Types of Databases

In discussions some scientists felt that the most important question they need to answer by querying databases is: "What is known and where can I find out more about it?" These investigators felt that databases containing only this information will be successful and that, regardless of other content, databases lacking this element will fail. They also expressed a sense of urgency, stating that because of the rapid rate at which data are being generated, at least some elements of a database or databases must be implemented within two years.

Some queries posed by scientists are now successfully answered by existing databases such as GenBank, PIR, and GRIN. Existing and planned databases can be modified to answer additional query types. However, some scientists have unique unmet needs that will require new developmental efforts in database design. These needs include electronic databases for stock centers or laboratories that provide seed stock; statistically annotated descriptions of phenotypic and environmental interactions and of quantitative traits affecting important agricultural characteristics; and horizontal integration of information between species for gene products and phenotypic characteristics, particularly agriculturally important ones. Some elements of these databases will require genetic representations that are unique to plants and original as database

descriptions, e.g., representing monosomic and trisomic lines; triploid tissues and polyploid cultivars; plant-specific cytogenetic features; somatic-cell ancestry. These are serious challenges that will require creative planning and integration.

CURRENT AND POTENTIAL TECHNOLOGICAL DEVELOPMENTS

The Plant Genome Research Program can and will benefit from new and emerging technological developments. This section reviews and lists some of the current research that could lead to breakthroughs in the collection, analysis and management of genome data. The list is not exhaustive; however, it represents some of the areas in which work is being done.

Database management

Cinkosky (1989) has reported on database changes at GenBank that portend improvements in system response and performance.

Merging and manipulation of data

Clark *et al.* (1990) of the Biomedical Computing Unit, Imperial Cancer Research Fund Laboratories, London are developing a computer-based system for integration and analysis of data and other information including sequences, secondary structures, x-ray studies, etc. from several sources. The use is for protein sequence analysis and other analyses, and will permit experimentation on hypothesis formation and other aspects of research. The system is called PAPAIN.

The National Center for Biotechnology Information has been established at the National Library of Medicine. One of its major goals is to facilitate electronic access in an integrated fashion to the many molecular biology databases (Benson *et al.* 1990). NLM is working on ways to express such information in a machine environment, and is also working on a knowledge base that can represent the metabolic map of *E. coli*.

Exchange of data

White and Allkin (1989) have reported efforts in Great Britain to develop an Exchange Data Format, (XDF) that will enable the ready exchange of data from various systems and on different topics in biology. At this point, it is most useful for the exchange of numeric, structured textual, and descriptive data.

Sequence analysis

Soderlund *et al.*, (1990)of the Computing Research Laboratory at New Mexico State University, developed a system called "gm" that automates the identification of genes in DNA sequence data. Conventional pattern analysis methods are used to identify components of genes. The system is in use at several laboratories.

MacInnes *et al.* (1990) of the Los Alamos National Laboratory and Bradley University, developed GENEX, a knowledge-based system for partially automated search and analysis of DNA sequences. Components of the system include analytic tools developed at the Genetics Computer Group, University of Wisconsin.

Overbeek (1990) has described work at Argonne National Lab on developing an expert system for machine reading of DNA sequence gels.

Shepp (1990) has described work at Bell Laboratories on automating DNA sequence analysis, particularly the aspects of automatic reading of gels and/or film.

Computerizing stock center databases

The *E. coli* Genetic Stock Center at Yale has implemented a relational database describing strains, mutations, and genes of *E. coli* (Berlyn and Letovsky personal communication) allowing complex queries against genotypes, mutations, structural mutations, and all other database objects. *Bacillus subtilis*, *Salmonella typhimurium*, and *Coenorhabditis* stock centers have computerized databases. The GRIN database, described in Appendix G, serves plant germplasm resources.

Genetic map databases

Letovsky and Berlyn, (1990) of Carnegie-Mellon University and Yale University respectively report work in progress on automated support for constructing and analyzing genetic maps at the *E. coli* Genetic Stock Center at Yale.

RFLP map software has been independently developed by Burr, Hoisington, and Coe. See Appendix B for further information.

Other research efforts

As reported in its research prospectus (1989), the University of Arizona's Computer and Genomic Systems Laboratory has experience on projects related to the electronic manipulation of genome data, and is seeking additional work in that arena.

References

Benson, D., M.S. Boguski, D. J. Lipman and J. Ostell. 1990. The National Center for Biotechnology Information. *Genomics* 6:389-391.

Cinkosky, Michael. The GenBank Databank reorganization. Presentation at the Conference on the matrix of biological knowledge (Bio-matrix 1989), Waterville Valley, NH, August 18-20, 1989.

Clark, Dominic A., Christopher J. Rawlings, Geoffrey J. Baron and Iain Archer. 1990. Knowledge-based orchestration of protein sequence analysis and knowledge acquisition for protein structure prediction. In *Artificial intelligence and molecular biology: Working notes. Proceedings of a symposium at Stanford University, March 27-29, 1990*, by the American Association for Artificial Intelligence. Menlo Park, CA: AAAI, 28-32.

Karp, Peter D. Hypothesis formation and qualitative reasoning in molecular biology. In *Artificial intelligence and molecular biology: Working notes. Proceedings of a symposium at Stanford University, March 27-29, 1990*, by the American Association for Artificial Intelligence. Menlo Park, CA: AAAI, 60-63.

Letovsky, Stanley, and Mary Berlyn. Constraint propagation techniques for genetic mapping. In *Artificial intelligence and molecular biology: Working notes. Proceedings of a symposium at Stanford University, March 27-29, 1990*, by the American Association for Artificial Intelligence. Menlo Park, CA: AAAI, 73-77.

MacInnes, Mark A., Sajeev Batra, Susan Roachl, and Gary C. Salzmanl. GENEX: A CLIPS knowledge-based program for automated search and analysis of molecular biological data bases. In *Artificial intelligence and molecular biology: Working notes. Proceedings of a symposium at Stanford University, March 27-29, 1990*, by the American Association for Artificial Intelligence. Menlo Park, CA: AAAI, 78-81.

Overbeek, Ross. "Automated interpretation of sequencing gels." AAAI: Menlo Park, CA, 1990. Paper presented at (but not included in the proceedings of) Artificial Intelligence and Molecular Biology, a symposium at Stanford University, March 27-29, 1990, by the American Association for Artificial Intelligence.

Research prospectus. November 1989. Tucson, AZ: Computer and Genomic Systems Laboratory, University of Arizona, 4 pages.

Shepp, Larry. "Mapping and interpreting biological information." Presentation at the workshop entitled Computing and Molecular Biology, Washington, D.C., April 30 - May 1, 1990, by the Computer Science and Technology Board of the National Research Council.

Soderlund, C. A., P. Shanmugam, and C. A. Fields. Integration of pattern analysis and geometric modeling in gm, an automated DNA sequence analysis system. In *Artificial intelligence and molecular biology: Working notes. Proceedings of a symposium sponsored by the American Association for Artificial Intelligence, Stanford University, March 27-29, 1990.* Menlo Park, CA: AAAI, 136-140.

White, R.J., and R. Allkin. "A language for the definition and exchange of biological data sets." Presentation at Bio-matrix 1989, a conference on the matrix of biological knowledge, Waterville Valley, NH, August 18-20, 1989.

APPENDIX A

USDA PLANT GENOME RESEARCH PROGRAM
SCIENCE AND TECHNOLOGY COORDINATING COMMITTEE
DATABASE SUBCOMMITTEE

Members:

Kevin Allred
Native Plant, Inc.
417 Wakara Way
Salt Lake City, Utah 84108
Phone: (801) 583-3500

Dennis A. Benson
Chief, Information Resources
National Center for Biotechnology Information
National Library of Medicine, NIH
Bethesda, Maryland 20209
Phone: (301) 496-2475
FAX: (301) 480-9241
INTERNET: DAB@NCBI.NLM.NIH.GOV

David Benton
GenBank/IntelliGenetics
700 East El Camino Real
Mountain View, California 94040
Phone: (415) 962-7360
FAX: (415) 962-7302
BITNET: BENTON@GENBANK.IG.COM
INTERNET: BENTON@KARYON.BIO.NET

Mary Berlyn
Department of Biology, Room 255, OML
Yale University, P. O. Box 6666
New Haven, Connecticut 06511
Phone: (203) 432-3536
FAX: (203) 432-3879
BITNET: BERLYN@YALEMED

Phil Filner
Correlation Genetics
2050 Concourse Drive
San Jose, California 95131
Phone: (408) 433-9808

Steve Heller
U.S. Department of Agriculture
Agricultural Research Service, BA
Room 164, Room 011A, BARC-West
Beltsville, Maryland 20705
Phone: (703) 442-0900

Robert Robbins
National Science Foundation
1800 G Street, N.W.
Room 215
Washington, D.C. 20550
Phone: (202) 357-7475
FAX: (202) 357-7745
INTERNET:
RROBBINS@NOTE.NSF.GOV
BITNET: RROBBIN@NSF

Keith Russell (Chair)
U.S. Department of Agriculture
National Agricultural Library, Rm. 100
Beltsville, Maryland 20705
Phone: (301) 344-3834
FAX: (301) 344-5472
BITNET: KRUSSELL@UMDARS
DIALCOM: 57:AGS3080
TELEMAIL: KWRUSSELL

Marvin Stodolsky
Health Effects Research Division
Office of Health & Environmental Research
U.S. Department of Energy
ER-72-GTN
Washington, D.C. 20545
Phone: (202) 353-3683
FAX: (202) 353-3884

Observers:

Machi Dilworth
National Science Foundation
1800 G Street, N.W.
Room 215
Washington, D.C. 20550
Phone: (202) 357-7475
FAX: (202) 357-7745
INTERNET:
MDILWORT@NOTE.NSF.GOV
BITNET: MDILWORT@NSF

Rose Broome
Computer Systems Analyst
Information Systems Division
5th Floor
National Agricultural Library
10301 Baltimore Boulevard
Beltsville, Maryland 20705
Phone: (301) 344-3813

Susan Fugate
Technical Information Specialist
Room 205
National Agricultural Library
10301 Baltimore Boulevard
Beltsville, Maryland 20705
Phone: (301) 344-3779
BITNET: SFUGATE@UMDARS

Leslie A. Kulp
Acting Head, Special Programs Branch
Room 1402
National Agricultural Library
10301 Baltimore Boulevard
Beltsville, Maryland 20705
Phone: (301) 344-3875

Janice Kemp
Technical Information Specialist
Room 1402
National Agricultural Library
10301 Baltimore Boulevard
Beltsville, Maryland 20705
Phone: (301) 344-3875
BITNET: JKEMP@UMDARS

Jean Larson
Technical Information Specialist
Room 304
National Agricultural Library
10301 Baltimore Boulevard
Beltsville, Maryland 20705
Phone: (301) 344-3704

Jerome P. Miksche
Director, Plant Genome Mapping Program
U.S. Department of Agriculture
Room 233, Building 005, BARC-West
Beltsville, Maryland 20705
Phone: (301) 344-2029

Quinn Sinnott
Germplasm Resources Information Network
U.S. Department of Agriculture
Room 128, Building 001, BARC-West
Beltsville, Maryland 20705
Phone: (301) 344-3072
FAX: (301) 344-3036

Eli Yecheskel
General Management Advisory Unit, Inc.
13825 Bethpage Lane
Silver Spring, Maryland 20906
Phone: (301) 460-7722

NAL staff who provided significant additional support for the work of the database subcommittee: Jannette Shuford-Hall, Rebecca Thompson, and Sandy Facinoli.

APPENDIX B

CORN DATASETS AND THEIR USES

Datasets and databases are more developed for corn than for other crop species. This appendix lists some of these and identifies specific users, types of queries, and perceived needs. It is not a complete listing; additions should be solicited from the corn research community.

I. CURRENT DATASETS AND DATABASES FOR CORN

A. *The Mutants of Maize*, by M. G. Neuffer (1968) contains the linkage maps, pictures of phenotypes organized on chromosome order, an index of genes and their chromosomal positions, and tables showing interactions of genes in pigment pathways.

B. The *Maize Genetics Cooperation News Letter* is an annual publication that provides information in the area of maize genetics, including notes, data compilation, nomenclature, 600-700 selected bibliographic references, and the addresses of scientists and labs. It also publishes:

1. *The Maize Genetic Coop Stock Center Catalog of Stocks* ordered by chromosome for one of the markers.

   Volume of use:
   3349 seed samples sent in response to 215 requests in 1988.
   2270 seed samples sent in response to 184 requests in 1987.

2. Linkage Maps: Linkage of markers and the RFLP map are presented together. Organellar genomes maps are included as well as references for linkage data.

C. The RFLP linkage databases were independently developed by Burr (Brookhaven National Laboratory, BNL); Hoisington (Centro Internacional de Mejoramiento de Maiz y Trigo, CIMMYT) and Coe *et al.* (University of Missouri); and Helentjaris and Wright (Native Plants Incorporated, NPI, Salt Lake City, Utah). There are efforts to integrate these databases.

The Burr Database is made available to investigators in the following way. Investigators get a set of recombinant inbreds; they probe and send data to Burr at BNL for analysis using a Fortran program on Vax hardware. Data are in the form of an array where columns are Parent 1 and Parent 2 alleles of the loci in the analysis and each row represents a recombinant inbred plant.

The program analyzes the polymorphisms and indicates to which columns the test alleles are related. The output is given in linear map order. The Burr lab has more than 40 lines in each of two separate families. Burr also has on an IBM PC an RBase database of genes and map position.

D. Databases of mutants and pedigrees (personal interview with M. G. Neuffer, University of Missouri). In addition to *The Mutants of Maize*, Neuffer has a database on mutants and phenotype, and on strains carrying these mutants. Neuffer also has a pedigree database. His databases are on an OS9 system Helix with 12 terminals. He characterizes them as fairly idiosyncratic. He would like to have all of these on an IBM or Mac compatible system to allow distribution.

II. CURRENT QUERYING OF THESE DATABASES:

A. Neuffer's strains database (see I.D above) receives well defined queries from researchers who usually have a specific trait, phenotype, or type of mutation, in mind and, at best, a guess about what the mutant phenotype might be. They wish to screen members of the mutant collection. He surveys the mutant collection and provides them with mutants of the postulated phenotype; they screen the selected ones. Below are four examples of such queries and postulated phenotypes:

1. Phytic acid mutants. The investigator thinks that they may result in defective kernel mutants because of differential distribution between endosperm and the rest of the seed. The investigator wants to delete kernel-defective mutants from the collection.

2. Photosynthetic electron transport across membrane. Several types of phenotypes were postulated. Some may look normal, but die at certain seedling age; some show specific effects of herbicides that affect electron transport; some have characteristic UV-fluorescent phenotype. The mutant collection is surveyed for these phenotypes.

3. Auxotrophs. The investigator postulates duplicate loci causing semilethal kernel traits. Semilethals are screened for response to supplementation.

4. Herbicide resistance. The investigator wants a large group of essentially unselected mutants and uses the specific herbicide to screen them.

B. BNL RFLP database. As previously described, Burr sends his set of recombinant inbreds to any requestor. They do blots & correlations with their trait of interest, send the data to Burr, who runs it through his program, sends back printout indicating linkage, with markers ordered by map position.

C. NPI RFLP database. NPI will map cloned genes sent by public investigators and send back the location of the gene. For research purposes only NPI can supply, upon request, subsets of clones to public investigators with corresponding data sheets of map location and patterns of five inbreds with three enzymes.

III. PERCEIVED NEEDS FOR UPGRADING THE EXISTING DATABASES.

The plant information resources we surveyed are not in relational format nor are they available on-line. The investigators would like to make the databases available on-line or alternatively make them available for distribution in IBM and MacIntosh formats. Although from these interviews there were no spontaneous suggestions to integrate the various types of databases, those interviewed endorsed the usefullness of integrating genetic and RFLP mapping information. Some maize researchers cited the need to have RFLP programs available for direct use, rather than channeling through database sources such as BNL.

References

Neuffer, M. G. 1968. *The Mutants of Maize*. Madison, Wisconsin: Crop Science Society of America.

Weber, David and Tim Helentjaris. 1989. Mapping RFLP loci in maize using B-A translocations. *Genetics* 121(3):583-590.

APPENDIX C

BACKGROUND INFORMATION ON CURRENT DATA SOURCES

This appendix provides information about databases that list molecular biology information sources, including genome databases; nucleotide sequence databases; and protein sequence and structure databases. Most of the information was taken from selected fields in the Directory of Biotechnology Information Resources (DBIR) and the Listing of Molecular Biology Databases (LiMB). These are the first databases included in this appendix. For each database in which other information sources were used, the sources are cited in the Publications list for that database.

Directory of Biotechnology Information/Resources

Other Names: DBIR

Database Description:

DBIR is a centralized directory to international sources of publicly available biotechnology information. It includes resources such as computerized databases and their distributors, networks, electronic bulletin boards, and other biological computer resources established for communicating and disseminating biotechnology data; culture collections and specimen banks; biotechnology centers; publications focusing on general issues in biotechnology; and nomenclature committees working on issues of nomenclature in biotechnology and molecular biology. Resource identification data include information such as names of organizations, publications, or databases, relevant addresses or phone numbers, and related DBIR records. Keywords, drawn from NLM's Medical Subject Headings (MESH) are assigned to each record and are searchable.

Source of data in records:

Staff create records based on brochures, articles and descriptive data sheets, and informative literature about resources. Verified by contributors.

Formats for accessing data in the database:

DBIR is maintained as an online file (DIRLINE) in MEDLARS on the National Library of Medicine's Toxicology Data Network (TOXNET). MEDLARS can be accessed with a personal computer and modem by direct dial or through the TELENET or TYMENET telecommunications networks. Floppy disks and paper copy are planned for future distribution.

Source of funding/support agency:

American Type Culture Collection and the National Library of Medicine.

Contact person for information about the database:

Kathleen Arnett, Information Specialist, Bioinformatics Department
American Type Culture Collection
12301 Parklawn Drive
Rockville, MD 20852
Phone: (301) 231-5527
BITNET: KA3@NIHCU; DIALCOM: 42:CDT0155

Listing of Molecular Biology Databases

Other Names: LiMB

Database Description:

LiMB contains information about molecular biology and related databases. The intent is to provide a coordinated effort to identify, characterize, and link databases relevant to molecular biology. Each entry contains 54 fields including: database staff names and addresses, database maintenance hardware and software; scope of coverage and database goals,; details of submission access to the data sets; database size; and types of data covered by the database. Begun as a simple directory, in the future LiMB will expand its role as a centralized information resource by providing a platform for designing front-ends for automatic access to distributed biological data sets.

Source of data in records:

LiMB entries are based on questionnaires filled out by the database managers or, in the absence of a completed questionnaire, on secondary sources, e.g., a journal article.

Update frequency: Annually

Formats in which data can be entered to the database:

Magnetic tape; floppy disk; electronic mail; paper.

Formats for accessing data in the database:

LiMB is available free of charge through one of the following media: electronic network mail, MS-DOS based floppy disk, or printed form.

Source of funding/support agency:

Los Alamos National Laboratory

Restrictions on acquiring/distributing data:

No limitations.

Contact person for information about the database:

Dr. John Lawton, LiMB Administrator
Theoretical Biology and Biophysics Group
Los Alamos National Laboratory
T10, MS K710
Los Alamos, NM 87545
Phone: (505) 667-9455
BITNET: LIMB@LANL.GOV

Publications describing the database:

Lawton, J. R., F. A. Martinez and C. Burks. 1989. Overview of the LiMB database. *Nucleic Acids Research* 17(15):5885-5899.

GenBank Nucleotide Sequence Data Bank

Other Names: GenBank

Database Description:

GenBank is responsible for collecting, organizing, maintaining, and distributing all published and an increasing number of unpublished nucleotide sequences of 50 bases and longer.

Source of data in records: Literature; direct submissions.

Update frequency:

Quarterly releases of distributed formats. GenBank Online (through IntelliGenetics) updated daily.

Formats in which data can be entered to the database:

Online; magnetic tape; floppy disk; electronic mail; paper.

Formats for accessing data in the database:

Online; magnetic tape; floppy disk; CD ROM; electronic mail; FTP; usenet newsgroup.

Source of funding/support agency:

National Institute of General Medical Sciences (NIGMS); National Library of Medicine (NLM); Department of Energy (DOE)

Restrictions on acquiring/distributing data:

No limitations

Contact person for information about the database:

Dr. David Benton
IntelliGenetics, Inc.
700 East El Camino Real
Mountain View, CA 94040
Phone: (415) 962-7360
BITNET: BENTON@GENBANK.1G.COM

Publications describing the database:

Burks, C. 1988. The GenBank database and the flow of sequence data for the human genome. *Basic Life Sciences* 46:51-56.

Moore, J. F., D. Benton, and C. Burks. 1989. The GenBank nucleic acid data bank. *Focus* 11(4): 69-72.

EMBL Data Library

Other Names:

European Molecular Biology Laboratory Nucleotide Sequence Data Library; EMBL Database; EMBL.

Database Description:

The European complement of GenBank, EMBL collects, organizes and distributes nucleic acid sequence data. The EMBL Nucleotide Sequence Data Library contains information on nucleotide sequences, including the actual sequences, functional features, index terms, literature citations, taxonomic identification, and other information. Data are collected and distributed in collaboration with GenBank, DDBJ, PIR, and SWISS-PROT.

Source of data in records:

Literature; direct submission; other databases (GenBank).

Update frequency:

CD-ROM and magnetic tape are released quarterly. Online format available immediately on entry with indices updated daily.

Formats in which data can be entered to the database:

Online; magnetic tape; electronic mail; paper.

Formats for accessing data in the database:

Online; magnetic tape; paper; CD-ROM; electronic mail.

Source of funding/support agency:
European Molecular Biology Laboratory

Restrictions on acquiring/distributing data:
There are no restrictions on the use or redistribution of the data.

Contact person for information about the database:
Dr. Graham N. Cameron
European Molecular Biology Laboratory
Postfach 10 22 09
Myerhofstrasse 1
6900 Heidelberg, Federal Republic of Germany
Phone: +49 6221 387258
FAX: +49 6221 387306
TELEX: 461613 (embl d)
EARN: DATALIB@EMBL

Publications describing the database:
Cameron, G. N., 1988. The EMBL data library. *Nucleic Acids Research* 16(5):1865-1866.

Kahn, P., and G. Cameron. The EMBL Data Library: a progress report. *Cytogenetics and Cell Genetics* 51(1-4):1020-1021.

Stoehr, P. J., and R. A. Omond. 1989. The EMBL network file server. *Nucleic Acids Research* 17(16):6763-6764.

DNA Data Bank of Japan

Other Names: DDBJ

Database Description:
Serving as the Asian complement to GenBank, DDBJ specializes in collecting nucleotide sequences generated in Japan and the surrounding region. Structure, content, and format are similar to that of GenBank and EMBL. DDBJ has close collaborative ties with GenBank and EMBL.

Source of data in records: Literature.

Update frequency: Semiannually

Formats in which data can be entered to the database:
Online; magnetic tape; floppy disk; hardcopy

Formats for accessing data in the database:
Online (accessible by direct-dial); magnetic tape; floppy disk; hardcopy

Source of funding/support agency:
Japanese government

Restrictions on acquiring/distributing data:
No limitations

Contact person for information about the database:
Dr. Sanzo Miyazawa, Manager
DNA Data Bank of Japan
Laboratory of Genetic Information Analysis
National Institute of Genetics
Mishima, Shizuoka, 411
Japan
Phone: (011-81) 559-75-0771 ext. 649
E-MAIL: SMIYAZAW%NIGSUN.NIG.JUNET@RELAY.CS.NET
DDBJ%NIGSUN.NIG.JUNET@RELAY.CS.NET

National Biomedical Research Foundation-Protein Identification Resource Protein Sequence Database

Other Names:
NBRF-PIR Protein Sequence Database; PIR Database; Dayhoff Database; NBRF Database

Database Description:
The NBRF-PIR Protein Sequence Database is one component of the Protein Identification Resource. This database includes the following: all substantially sequenced proteins, including sequences translated from nucleic acid sequences; bibliographic citations for amino acid sequences, nucleic acid sequences, X-ray crystallography, active site determination, etc.; annotations identifying posttranslational modifications, active sites, signal sequences, activation cleavages, disulfide bonds, intron locations, etc. An auxiliary file includes sequences in preparation and fragmentary and "hypothetical" sequences. Annotations show the locations of protein coding regions. This protein data bank serves as the U.S. collection node in collaboration with JIPIDS (Japan) and MIPS (FRG) under the association of PIR-International, their official coordinative body.

Source of data in records:
Literature (scientific journals and manuscripts) and direct submission.

Update frequency: Quarterly

Formats in which data can be entered to the database:
Online; magnetic tape; floppy disk; electronic mail; hard copy.

Formats for accessing data in the database:
Online (accessible through PROPHET, NBRF-PIR, MBIS, MBCRR, Biocomputing Research Unit in Molecular Biology, Pittsburgh Supercomputing Center); magnetic tape (available through NBRF-PIR and EMBL Data Library); staff search; floppy disc; compact disc.

Source of funding/support agency:
Division of Research Resources, National Institutes of Health (grant RR01821)

Restrictions on acquiring/distributing data:
No limitations.

Contact person for information about the database:
Winona Barker, Director
Protein Identification Resource (PIR)
National Biomedical Research Foundation (NBRF)
Georgetown University Medical Center
3900 Reservoir Road, N.W.
Washington, DC 20007
Phone: (202) 671-1662
BITNET: PIRMAIL@GUNBRF

Publications describing the database:
Sidman, K.E., D. G. George, W. C. Barker and L. T. Hunt. 1988. The protein identification resource. *Nucleic Acids Research* 16(5):1869-1871.

Martinsreid Institute for Protein Sequence Data

Other Names: MIPS

Database Description:
Serves as the European partner to PIR-International. The database contains bibliographic and protein sequence data. Data is collected and distributed in collaboration with JIPIDS and NBRF-PIR.

Source of funding/support agency:
European Economic Community

Restrictions on acquiring/distributing data:
No limitations.

Contact person for information about the database:
Dr. Hans-Werner Mewes
Martinsried Institute for Protein Sequence Data
Max Planck Institute for Biochemistry
AM Klopferspitz 18A
8033 Martinsreid
Federal Republic of Germany
BITNET: MEWES@DMOMPB51

International Protein Information Database in Japan, Sequence Database

Other Names: JIPIDS

Database Description:
JIPIDS contains nucleotide and amino acid sequences from Asian-based research. The following specific fields are covered: plants, plant viruses, T4 phage, amylases, ferredoxins, kalikreins, calmodulins, and thioredoxins. As a member of PIR-International, JIPIDS works cooperatively with MIPS and NBRF-PIR. It also collaborates closely with DDBJ.

Source of data in records: Literature (books and journals).

Update frequency: every 3-4 months

Formats in which data can be entered to the database:
Online; magnetic tape; floppy disk; electronic mail; hardcopy.

Formats for accessing data in the database:
Online; magnetic tape.

Source of funding/support agency:
Japanese government; research funds; donations.

Contact person for information about the database:

Akira Tsugita
JIPID
Life Science Institute
Science University of Tokyo
Yamazaki Noda 278
Japan
Phone: [011-81] 471-24-1501 ext. 5001 or [011-81] 471-22-3899
FAX: 0471-22-1544
BITNET: TSUGITA@JPNSUT31
EX5292@JPNSUT30
DIALCOM: 42:CDT0079

Publications describing the database:

Tsugita, A. 1989. Current status of protein data banks. In *Methods in Protein Sequence Analysis*, ed. B. Wittmann-Liebold. New York: Springer-Verlag.

SWISS-PROT Protein Sequence Data Bank

Other Names: SWISS-PROT

Database Description:

SWISS-PROT contains protein primary sequence data built by reformatting the data provided in the NBRF-PIR Protein Sequence Database and by translating the EMBL Nucleotide Sequence Data Library. Data are collected and distributed in collaboration with EMBL.

Update frequency: 3 to 4 times per year.

Formats for accessing data in the database:

Online (accessible through GenBank Online, MBIS, Pittsburgh Supercomputing Center); magnetic tape (available through EMBL Data Library); floppy disc; compact disc.

Contact person for information about the database:

Amos Bairoch
Departement de Biochimie Medicale
Centre Medical Universitaire
1, rue Michel-Servet
1211 Geneva 4
Switzerland
Phone: +41 22 468758
BITNET: BAIROCH@CGECMU51

Protein Data Bank

Other Names: PDB

Database Description:

PDB seeks to provide comprehensive coverage of bibliographic, atomic coordinate, and crystallographic structure factor data for biological macromolecules such as proteins, tRNAs, polynucleotides, viruses, and polysaccharides. Bibliography files provide bibliographic information about macromolecular structures for which coordinates are not yet available. Program files provide source codes for FORTRAN programs for accessing and manipulating the data files. Cross referenced to the Atlas of Protein Sequence and Structure (PIR).

Source of data in records:

Data are directly submitted by principal investigators or are from the literature.

Update frequency: Quarterly.

Formats in which data can be entered to the database:

Magnetic tape; floppy disk; electronic mail.

Formats for accessing data in the database:

Online (accessible through PROPHET, CAN/SND, EMBL Data Library, Pittsburgh Supercomputing Center); hardcopy; microfiche; magnetic tape.

Source of funding/support agency:

National Science Foundation; National Institutes of Health; Department of Energy; user fees.

Restrictions on acquiring/distributing data:

There are no restrictions on availability of data. It is improper for recipients to offer computer files, magnetic tapes, microfiche, listings, or any further Protein Data Bank product or copies for sale as a commercial item.

Contact person for information about the database:

Ms. Frances C. Bernstein, Computer Analyst
Protein Data Bank
Chemistry Department, Brookhaven National Laboratory (BNL)
Building 555
Upton, NY 11973
Phone: (516) 282-4382
FAX: (516) 282-3000, (516) 282-4315
BITNET: PDB@BNLCHEM

NEWAT Protein Data Base

Database Description:
Molecular biology database; factual; protein sequence data.

Formats for accessing data in the database:
Online (accessible through MBIS)

Contact person for information about the database:
R. F. Doolittle
University of California
Department of Chemistry
La Jolla, CA 92093

PROSEQ Protein Data Base

Database Description:
Database; factual; protein sequence data

Formats for accessing data in the database:
Online (accessible through MBIS)

Contact person for information about the database:
Dr. T. Ooi
Institute for Chemical Research
Kyoto University
Uji, Kyoto-Fu, 611, Japan

Cambridge Structural Database

Other Names: CSD

Database Description:
Provides bibliographic, chemical connectivity, and numeric data for X-ray and neutron diffraction studies of organo-carbon compounds (organics, organometallics, and metal complexes). Comprises three files: 1) BIB which contains the bibliographic citation for each diffraction study along with some textual chemical information, 2) CONN which provides a connectivity representation of the chemical structure and 3) DATA which provides the numeric results.

Source of data in records:
Primary journals, chemical abstracts, service journal editors, and scientists.

Update frequency: Quarterly

Formats for accessing data in the database:

Online (accessible through PROPHET, CAN/SND, FIZ, CSSR, Pittsburgh Supercomputing Center); magnetic tape (available through Cambridge Crystallographic Data Centre and National Affiliated Centres in thirty countries); staff search (available through the Medical Foundation of Buffalo to licensed users).

Restrictions on acquiring/distributing data:

Restricted to those countries accredited by the Cambridge Crystallographic Data Centre and to users with an access agreement license.

Contact person for information about the database:

Dr. Olga Kennard
Cambridge Crystallographic Data Centre
University Chemical Laboratory
Lensfield Road
Cambridge, United Kingdom
CB2 1EW
Phone: (0223) 336409
E-MAIL: OK10%CHEMCRYS@UK.AC.CAM
FAX: (-44) 223 336362
TELEX: 81240 CAMSPL G (telex)

APPENDIX D

PUBLIC AND PRIVATE SECTOR ORGANIZATIONS WITH RFLP CAPABILITIES

Note: A preliminary list was prepared by Mark Walton, Technical Applications Manager, Ceres (a division of Native Plants, Inc.). Additional names were suggested by subcommittee members and observers. Further work is needed to fully represent organizations conducting RFLP research.

A. PRIVATE SECTOR COMPANIES WITH RFLP CAPABILITIES

Company	Contact	Crop
Asgrow Seed Co.	Dr. John Sorensen 7000 Portage Rd. Kalamazoo, Michigan 49001 (616) 385-6699	Corn
Agrigenetics Corp.	Dr. Mike Murrary 5649 E. Buckeye Rd. Madison, Wisconsin 53716 (608) 221-5000	Corn
Agrimont	Dr. Carlo Minganti Centro de Biotechnologie Via Massa Avenza 85 54100 Massa, Italy	Corn, Alfalfa
Agricola Mais Ibridi	Dr. Daniele Gilberti Via Grazie, 6 25122 Brescia, Italy	Corn
Biogenetic Svcs	Dr. Alex Kahler 2308 6th Street E. P.O. Box 710 Brookings, South Dakota 57006 (605) 697-8500	Corn, Soybeans
Biosem/Limagrain	Dr. Joel Perret Lab. de Bioligie Cellular & Molecular 24, Avenue des Landais 63170 Aubiere, France	Corn

Company	Contact	Crop
CIBA-GEIGY LTD.	Dr. Phillipe Gay Agricultural Division CH-4002, Basle Switzerland	Corn
Dupont	Dr. Scott Tingey E402/4249 Experimental Station Wilmington, Delaware 19898 (302) 695-7252	Soybeans, Corn
Ed. J. Funks/BP	Dr. Sue Sullivan 601 Funk Pkwy Box 67 Kentland, Indiana 47951 (219) 474-5111	Corn
Garst Seed/ICI	Dr. Ian G. Bridges Research Department Highway 210, P.O. Box 500 Slater, Iowa 50244 (515) 685-3574	Corn
Hillenshog AB	Dr. Chris Bornman Box 302 S-261 23 Landskrona Sweden	Sugar Beets
KWS	Dr. J. F. Seitzer Girmsehlstrasse 31 D-3352 Einbeck 1 Federal Republic of Germany	Corn
Nickerson Seed	Dr. Iain Cubitt Cambridge Science Park Milton Road Cambridge CB4 4WE United Kingdom	Corn
Northrup King/Sandoz	Dr. Ed Weck Research Center Stanton, Minnesota 55081 (507) 645-5621	Corn

Company	Contact	Crop
ORSAN	Mrs J. Simon-Genilloud 16, Rue Ballu F 75009, Paris, France	Corn
PBI/Unilever	Dr. Peter Payne Maris Lane, Trumpington Cambridge CB2 2LQ United Kingdom	Wheat
Pioneer Hi-Bred	Dr. John Howard Department of Biotech 7300 N.W. 62nd Avenue Johnston, Iowa 50131 (515) 270-3650	Corn
Rhone-Poulenc	Dr. J. L. Arnault 14-20 Rue Pierre Baizet BP. 9163 69263 Lyon Cedex 09 France	Corn
United Agriseeds/Dow	Dr. Alan Gould P. O. Box 4011 Champaign, Illinois 61820 (217) 373-5300	Corn
D. J. Van der Have	Dr. Case Noome Plant Breeding Station 4410 AA Rilland P. O. Box 1 The Netherlands	Corn, Rape

B. PUBLIC RFLP RESEARCH PROGRAMS

Arabidopsis

Howard Goodman
Department of Molecular Biology
Massachusetts General Hospital
Harvard Medical School
Boston, Massachusetts 02114
(617) 726-5933

Elliot Myerwitz
Department of Biology, 156-29
California Institute of Technology
Pasadena, California 91125
(818) 356-6889

Christopher Somerville
DOE Plant Research Lab
Michigan State University
East Lansing, Michigan 48824
(517) 355-5159

Brassica

Tom Osborn
Department of Agronomy
University of Wisconsin
Madison, Wisconsin 53706
(608) 262-2330

Carlos Quiros
Department of Vegetable Crops
University of California, Davis
Davis, California 95616
(916) 752-1734

Barley

Tom Blake
Department of Plant and Soil Science
Montana State University
Bozeman, Montana 59717-0002
(406) 994-5055

Citrus

Mikeal Roose
Department of Botany and Plant Science
University of California - Riverside
Riverside, California 92521

Corn

Paul Sisco
Department of Genetics
North Carolina State University
Raleigh, North Carolina 27695-7614
(919) 737-2704

David Hoisington
CIMMYT
Apdo Postal 6-641, 06600, DF
Mexico, Mexico

Jack Gardiner
Department of Agronomy
302 Curtis Hall
University of Missouri
Columbia, Missouri 65211
(314) 875-5359

Mike Lee
Department of Agronomy
Iowa State University
Ames, Iowa 50011
(515) 294-3052

Ben Burr
Brookhaven National Lab
c/o Biology Department
Building 463
Upton, NY 11973
(516) 282-3396

Corn, continued

Charles Stuber
Department of Genetics
North Carolina State University
Box 7614
Raleigh, NC 27695
(919) 737-2289

Cotton

Glen Galau
Botany Department
University of Georgia
Athens, Georgia 30602

Dennis Ray
Department of Plant Science
University of Arizona
Tucson, Arizona 85721

David Stelly
Texas A&M University
Soil & Crop Sciences Department
College Station, TX 77843
(409) 845-2745

Lettuce

Richard Michelmore
Department of Vegetable Crops
University of California - Davis
Davis, California 95616
(916) 752-1729

Pepper

Steve Tanksley
Department of Plant Breeding and
Biometery
Cornell University
Ithaca, New York 14853
(607) 255-7886

Poplar

Elizabeth Van Volkenburg
Department of Botany, KB-15
University of Washington
Seattle, Washington 98195
(206) 543-1942

Potato

Steve Tanksley
Department of Plant Breeding and
Biometery
Cornell University
Ithaca, New York 14853
(607) 255-7886

Christine Gebhart
Max Planck Institut fur
Zuchtungsforschung
D-5000 Koln 30
Max Planck Institute
Federal Republic of Germany

Rice

Steve Tanksley
Department of Plant Breeding and
Biometery
Cornell University
Ithaca, New York 14853
(607) 255-7886

Sorghum

Gary Hart
Department of Soil and Crop Science
Texas A&M University
College Station, Texas 77843
(409) 845-8293

Soybeans

Randy Schumacher
Department of Agronomy
Iowa State University
Ames, Iowa 50011
(515) 294-6233

Gordon Lark
Department of Biology
University of Utah
Salt Lake City, Utah 84108
(801) 581-6517

Wheat

Gary Hart
Department of Soil and Crop Science
Texas A&M University
College Station, Texas 77843
(409) 845-8293

Steve Tanksley
Department of Plant Breeding and Biometery
Cornell University
Ithaca, New York 14853
(607) 255-7886

Bikram Gill
Department of Plant Pathology
Throckmorton Hall, Room 414
Kansas State University
Manhattan, KS 66506
(913) 532-6176

APPENDIX E

A NATIONAL PLANT GERMPLASM COMMITTEE SPECIAL REPORT: GENETIC STOCK COLLECTIONS

United States Department of Agriculture | Agricultural Research Service | Office of the Administrator | Washington, D.C. 20250

PH: 317-494-6573
FTS-284-6573

Agricultural Science Advisor
Plant Germplasm

Agronomy Department
Rm.202, Poultry Science Bldg.
Purdue University
W. Lafayette, IN 47907

August 16, 1989

SUBJECT: National Plant Germplasm Committee
Special Report--Genetic Stock Collections

TO: Members of the Plant Germplasm Community

FROM: Paul J. Fitzgerald, Chairman, NPGC

A joint meeting of the National Plant Germplasm Committee and the Curators of the major genetic stocks collections was held at Urbana/Champaign, Illinois on May 1-3, 1989 to review practices and procedures and recommended policy on the future management of these collections.

The attached report was developed by a subcommittee of the NGPC appointed to hear the Curator reports,analyze the information, and recommend action by appropriate components of the National Plant Germplasm System to insure the continuing support and management of these genetic stocks to maintain their integrity and long term security.

This report has been or will be provided to members of the NPGC, the National Plant Genetic Resources Board, the ARS Germplasm Matrix Team, Curators of Genetic Stock Collections, participating researchers and Administrators.

A National Plant Germplasm Committee Special Report[1]

GENETIC STOCK COLLECTIONS
Executive Summary

The National Plant Germplasm System (NPGS) is a coordinated network of institutions, agencies, and research units representing Federal, State, and Industry sectors, working cooperatively to introduce, maintain, evaluate, enhance, catalog, and distribute plant germplasm. It is a user oriented system with a goal of providing to the user community sufficient genetic diversity for crop improvement.

Peripheral to the central repositories of the NPGS are several genetic stock collections which receive part State and part Federal funding. Most are maintained by a single state or federal scientist with inadequate funding. The procedures for saving these collections when the scientists retire or leave for other reasons are generally nonexistent. In some cases 50 years were required to assemble the collections. It is essential that these collections be evaluated and approximately saved.

Nine species with generic stock collections (maize, wheat, crucifer, sorghum, cotton, barley, pea, tomato, and soybean) have been identified. The collections contain from a few hundred to 80,000 accessions. There are induced and natural mutations, aneuploids, chromosomal aberrations, genetically engineered material, and wild species.

The genetic stock collections have always been important to plant scientists. With the advent of biotechnology and genetic engineering they are becoming much more important. Contrasting traits are essential for studies designed to determine gene action and gene location on chromosomes. Genetic research and molecular research such as RFLP analyses require the genetic variation represented in these genetic collections. Such collections are necessary for graduate programs which train germplasm curators. Maintaining these collections is essential for continued plant improvement.

Recommendations:

1. Begin immediately the process for obtaining agreement from curators and their administrators that collections will be maintained or transferred with full documentation to another location should the curator resign, retire, etc.

---

[1] From a Review held May 1-2, 1989 at Urbana, Illinois by the National Plant Germplasm Committee

2. Begin as soon as is feasible the process of building a genetic stock collection database within GRIN. The database should be such that all necessary information is included in an accessible format. It should not; however, be an integral part of the GRIN germplasm database.

3. Develop plans for base funding of the genetic stock collections. The plans may, or may not, require that curators be ARS employees. Each case may be different and the memo of understanding to be negotiated will determine the direction to take.

4. Survey all collections of genetic material to determine which qualify as "Genetic Stock Collections." Before this step can be taken, criteria for such collections must be developed.

5. A subsample (25-50 seeds) of each genetic stock should be put into cryogenic storage at NSSL. Curators can determine what material must be included so that all genotypes, aneuploids, aberrations, wild species, etc. are represented. Adequate descriptions must be submitted to NSSL for each sample.

6. Include, as part of the genetic stock collection funding, funds for training graduate students to become germplasm and genetic collection curators.

July 28, 1989

Merle H. Niehaus, Chair, Colorado State University
Norman James, ARS/USDA
Larry Baker, Asgrow Seed Company

A National Plant Germplasm Committee Special Report [1]

GENETIC STOCK COLLECTIONS
July 28, 1989

The National Plant Germplasm System (NPGS) encompasses both state and federal programs and is a structured organization which addresses the issues having to do with acquisition, preservation, enhancement, distribution, and utilization of plant germplasm. It includes four regional (W-6, NC-7, S-9, and NE-9) and one interregional (IR-1) Plant Introduction Stations, several national clonal repositories, germplasm services units at Beltsville, MD, and the National Seed Storage Laboratory (NSSL) at Fort Collins, Colorado. Among the germplasm services are plant exploration, plant introduction, plant quarantine, and the Germplasm Resources Information Network (GRIN) database, the centralized database of the NPGS. A combination of State Agricultural Experiment Station (SAES) state funds, SAES Regional Research funds (federal), and Agricultural Research Service (ARS) funds are the funding sources for the Plant Introduction stations.

Because virtually all the major crops grown in the U.S. originated elsewhere, genes occurring in the native habitat in landraces or wild species must be collected and saved if they are to be available for future crop improvement. Many are in the NPGS now and are being preserved. Additions to the collections are necessary because the native habitats for many species are undergoing agricultural or urban development causing the disappearance of the landraces and wild species.

Peripheral to the central repository activities of the germplasm system are a number of genetic stock collections. Some are at university sites and some are at ARS sites. Most are managed by dedicated scientists who have developed the collections during their careers. Some have formal funding and some are part of a plant scientist's project with no dedicated formal funding. some have been recognized as being critical collections and do receive ARS funds. However, in several instances, the ARS funds are permanently budgeted items and support can vary from year to year. Several of the collections also utilize funds from private industry.

The genetic stock collections are variable in composition, but include stocks carrying natural and induced point mutations, aneuploids, chromosome aberrations of many types, wild species, genetically engineered material, and other accessions of potential interest to plant scientists. Often there is no other documented source of the material.

---

[1] From a Review held May 1-2, 1989 at Urbana, Illinois by the National Plant Germplasm Committee

Currently the recognized genetic stock collections include maize, wheat, crucifer, sorghum, cotton, barley, pea, tomato, and soybean, There are probably others which qualify as genetic stock collections and there are probably species not having such a collection where one is needed. In all cases the curators of the collections provide material to any plant scientist who requests it. No definitive criteria have been developed for determining whether a collection qualifies to be a genetic stock collection, and there are no published guidelines regarding genetic stock collection management.

Several problems exist, but there is a consensus that the genetic stock collections should be made a more integral part of the NPGS. Problems needing attention are: funding mechanisms, responsibility for the collections, and criteria for determining whether a collection qualifies.

Funding and responsibility are interconnected. Collections at state universities receive most of their funding from the state. Therefore, the responsibility for the collection is at the state level. With budgets at state institutions under pressure, and with a need to reallocate funds to high priority areas such as biotechnology, the collections are a risk of being lost. Ideally, another source of funding would be made available and the collections could be maintained without interruption even if state funding were withdrawn. The only obvious source of such funding is from ARS, which takes the national leadership germplasm preservation.

If ARS funding is not immediately available, curators should insure that collections are not discarded if state funding is withdrawn. A first step is to make it clear to curators who are near retirement, that they should make arrangements for storing, transferring, or distributing the collection if no new curator is being considered. This should be initiated well before retirement and should certainly include documentation of the material. NPGS coordinators should be advised before the beginning of the final year. Perhaps more important, the administrators in charge of programs where collections are located must agree to preserve the collections in case of death, disability, unexpected retirement, or resignation of the curator. Such agreement should be formalized with a memo of understanding or other formal agreement.

Recommendations follow:

1. Begin immediately the process for obtaining agreement from curators and their administrators that collections will be maintained or transferred with full documentation to another location should the curator resign, retire, etc.

2. Begin as soon as is feasible the process of building a genetic stock collection database with GRIN. The database should be such that all necessary information is included in an accessible format. It should not, however, be an integral part of the GRIN germplasm database.

3. Develop plans for base funding of the genetic stock collections. The plans may, or may not, require that curators be ARS employees. Each case may be different and the memo of understanding to be negotiated will determine the direction to take.

4. Survey all collections of genetic material to determine which qualify as "Genetic Stock Collections." Before this step can be taken, criteria for such collections must be developed.

5. A subsample (25-50) seeds of each genetic stock would be put into cryogenic storage at NSSL. Curators can determine what material must be included so that all genotypes, aneuploids, aberrations, wild species, etc. are represented. Adequate descriptions must be submitted to NSSL for each sample.

6. Include as part of the genetic stock collection funding, funds for training graduate students to become germplasm and genetic collection curators.

Brief descriptions of most of the genetic stock collections now being maintained follow:

1. Barley

   The barley collection is headquartered in the Agronomy Department of Colorado State University at Fort Collins, Colorado. The curator is Dr. Takumi Tsuchiya, a state employee in the department. As well as maintaining the collection, Dr. Tsuchiya and others publish a Barley Genetics Newsletter.

   The group responsible for the collection is international and individuals from various parts of the world have agreed to be responsible for various parts of the collection. Coordinators for maintaining the seven linkage groups are from the U.S., Canada, Japan, and Denmark. The genetic stocks are maintained in Fort Collins, Colorado and in Germany, Sweden, Denmark, and Canada.

   A handbook has been developed which includes instruction on how to grow the stocks so that the genes or aberrations are maintained. The scientists working with barley are constantly adding new material to the collections.

   For several years the Fort Collins work was funded by the state and by the National Science Foundation. The NSF funding is no longer available and the state with help from ARS is now providing the funding. State funding is not dependable because the state legislature has been decreasing its support for agricultural programs each of the past 10 years. Until this year there have not been budget cuts; there have been major shortfalls because funding did not cover costs. This collection is currently being regrown to provide fresh samples for cryopreservation at NSSL.

Several of the barley curators are near retirement. Dr. Tsuchiya has announced that he will retire within four years. A policy must be developed and implemented soon if the barley genetic stock collections are to be maintained.

2. Tomato

The tomato stock collection is held entirely at the University of California at Davis, California. Base collection samples of the tomato genetic stock collection have been sent to the NSSL. The curator is Dr. Charles Rick. The collection has developed over a 45 year period under the guidance of Dr. Rick. The collection receives state and ARS support. Like the barley collection, NSF provided support for several years but has not continued that support.

The collection includes about 1,000 accessions of wild species, 700 monogenic stocks, 800 miscellaneous stocks, and 200 unassimilated stocks. The material is used widely in the U.S. and around the world. Resistance to at least 32 major tomato diseases has been discovered in tomato exotics and of these, 16 resistances have been bred into commercial tomato cultivars. the tomato has become a model system for biotechnology research and the stock collection is vital for progress in biotechnology.

The tomato collection's curator is near retirement and the current funding is not adequate to continue the collection at its current size and activity. A policy for genetic stock collection must be developed and implemented soon to insure the integrity of the tomato genetic stock collection.

3. Pea

The pea genetic stock collection is held at the New York Agricultural Experiment Station at Geneva, New York. The curator who developed the collection, Dr. Gerry Marx, died in November, 1988 and the future of the collection is being evaluated. Dr. Stig Blixt of the Nordic Gene Bank is helping with that evaluation on a temporary basis as is Mr. Joseph Covert of Cornell University. Dr. Marx is not likely to be replaced, at least not by someone with the same interests and responsibilities. As is the case in other states, the New York Agricultural Experiment Station is under budget constraints.

There are 80,000 accessions in the collection.[1] The current physical facilities are adequate. A computerized database has been implemented.

[1] Breeding lines and many varied gene combinations are included.

A plan has been developed which outlines what will be required over the next five years to maintain the collection. Under the plan the 80,000 accessions would be reduced to only 500 stocks over the next five years.[2]

The pea collection is highly vulnerable now and a policy must be developed and implemented very soon in order to maintain the collection. Like other states, the New York Agricultural Experiment Station is under budget constraints.

4. Soybean

The soybean collection is held in an ARS laboratory in conjunction with the University of Illinois at Urbana, Illinois. The current curator is Dr. Randy Nelson. He was appointed in 1988, after Dr. Richard Bernard retired. Unlike several of the genetic stock collections, the soybean genetic stocks are not in a dedicated collection. They are an integral part of the soybean germplasm collection held at Urbana.

Included in the collection is a genetic type collection. this collection holds only unique material derived by mutation. The type collection contains 310 accessions with 151 stocks to be added. An isoline collection contains 300 accessions with 100 more to be added soon. Dr. Reid Palmer, an ARS Scientist at Iowa State University, Ames, Iowa, also maintains a soybean aneuploid genetic stock collection.

Funding does not appear to be a major problem specific to the soybean genetic stock collection. Perhaps a genetic stock soybean database should be developed and funds will be needed to do this. Apparently there are many accessions in the germplasm collection which might better be considered genetic stocks.

Separate from the ARS laboratory, Dr. Ted Hymowitz of the University of Illinois Agronomy Department maintains a collection of related wild species with funding from NPGS. This collection has 492 accessions representing 12 species.

5. Maize

The maize genetic stock collection is held in the Agronomy Department at the University of Illinois in Urbana, Illinois. The curator is Dr. Earl Patterson. The collection began at Cornell University in New York and was moved to Illinois in 1952. There are now 75,000 individually pedigreed seed samples. Chromosomal

---

2 The 500 lines represent a special group of lines developed by Dr. Marx. The Regional Plant Introduction Station in cooperation with the New York AES, Geneva and the Nordic Gene Bank will assume responsibility for the genetic stock collection. The Nordic Gene Bank maintains and distributes some 5,000 of the lines. The remainder, mostly breeding lines and combinatorial gene duplicates will be made available to interested breeders and geneticists worldwide through pea stocks newsletters.

aberrations make up 1,000 samples with 875 translocation stocks. From 75 to 100 new stocks are added each year.

Demand for the material has always been high, but it increased by 50% last year. About 25% of the distribution is to foreign countries and this figure appears to be true for the other collections as well. There are also many requests for information about the collection. A newsletter is published annually which includes a list of material as well as germplasm related research papers.

The maize collection should be put into a database as soon as possible. The database must be inclusive enough to include virtually all of the information needed. Dr. Patterson is near retirement and someone should be assigned the task of developing the database while Dr. Patterson is still there.

A policy which will fund the database and provide needed cold storage as well as one which will guarantee the integrity of the collection should be implemented.

6. Wheat

The wheat genetic stock collection is held at the University of Missouri at Columbia, Missouri. The curator is Dr. Ernie Sears. The collection has focused on the development and the maintenance of nullisomics, monosomics, trisomics, tetrasomics, telocentrics, mono-telocentrics, di-telocentrics, etc. These stocks can be used to locate genes on chromosomes, locate genes on arms of chromosomes, to measure distance between genes and the centromere, and to identify aneuploids.

Wheat is a hexaploid and mutations are difficult to detect. Therefore, there are few mutants in the collection. Much of the work is aimed at developing and maintaining the chromosome stocks. In most cases there are complete sets of the various aneuploids.

The aneuploids portion of the collection includes about 350 different aneuploids, all in a single cultivar, Chinese Spring. There are 275 species and varieties of wheat and its relative, wheat mutants, induced amphiploids, intervarietal chromosome substitutions, and lines with introduced segments of alien chromosomes.

Most of the aneuploids in the collection are duplicated at the Institute for Plant Science Research at Cambridge, England and at the University of Sydney in Australia. Samples of nearly all the material are in cryogenic storage in NSSL.

7. Sorghum

The sorghum genetic stock collection is held at the ARS Southern Crops Research Laboratory, Crop Germplasm Research in cooperation with the Texas Agricultural Experiment Station at College Station, Texas. The collection is not identified separately from the research program on sorghum. The existence and maintenance

of the collection is a result of the individual researcher's needs and his personal commitment to the collection. The current curator is Dr. K. F. Schertz.

The collection contains accessions having 240 specific morphological traits, five male-sterility inducing cytoplasms, and 15 translocation stocks (total = 260 lines). If funding permits, molecular markers will be added. However, funding is barely adequate to continue at current levels of support. Dr. Schertz is near retirement and the continuation of the collection depends somewhat upon the person who replaces him. A modification of the research project's scope is needed which will insure the integrity of the sorghum collection. Under current policy it is not certain that the collection will be maintained after Dr. Schertz retires.

8. Cotton

The cotton genetic stock collection is held at the ARS Southern Crops Research Laboratory, Crop Germplasm Research in cooperation with the Texas Agricultural Experiment Station at College Station, Texas. The collection is not identified separately from the research program on cotton. The existence and maintenance of the collection is a result of individual researcher's needs and their personal commitment to the collection. The current curators are Dr. D. M. Stelly and Dr. R. J. Kohel. Dr. E. L. Turcotte maintains some stocks of *Gossypium barbadense* in Arizona.

The collection includes 350 accessions having morphological markers and testers, 45 monosomes and teleosomes, 124 translocations, and 12 chromosome substitution lines (total = 531 lines).

The research project's scope should be modified so that there is insurance that the cotton genetic stock collection will be maintained. Current policy would apparently allow the collection to be discarded if the scientist in charge were not interested in maintaining it.

9. Crucifer

The crucifer genetic stock collection is maintained by Dr. Paul H. Williams, Department of Plant Pathology, University of Wisconsin at Madison, Wisconsin. It is the only collection included in this review not funded in full or in part by USDA funds. The predominant number of accessions in the collection belong to the genera *Brassica*, *Raphanus*, and *Arabidopsis*. Virtually all the material has been incorporated into genetic backgrounds resulting in extremely short reproductive cycles (6-10 cycles per year).

Dr. Williams has developed the Crucifer Genetics Cooperative as an organization dedicated to the distribution of seed and information on Cruciferae for the benefit of research and education. Currently over 1300 scientists representing 44 countries are members.

There are presently over 80 stocks of *Brassica* and 250 stocks of *Arabidopsis* available for distribution. More than 100 unique stocks are being developed and are nearly ready for distribution. Many more stocks are in early stages of development. The collection also includes pollen, gene libraries, and crucifer symbionts.

The Cooperative holds workshops every 18 months with between 100 and 250 attendees. A Resources Book is published which contains a list of materials, instructions on how to use the material for research and education, and other items of interest.

Dr. Williams has proposed that the collection be incorporated into the National Plant Germplasm System and that funding be provided. Any policy developed should address this proposal.

10. Miscellaneous

There are other collections which should be considered as genetic stock collections. An example is *Datura* where Dr. T. Tsuchiya at Colorado State University has a small collection. There are almost certainly other such collections around the country which should be identified and evaluated. Without a well considered policy which should be implemented soon, such collections are even more likely than the major collections to be discarded when the curator leaves.

Merle H. Niehaus, Chair, Colorado State University
Norman James, ARS/USDA
Larry Baker, Asgrow Seed Company

(Dr. Robert Plane, New York Agricultural Experiment Station and Dr. Richard Lower, Wisconsin Agricultural Experiment Station were also members of the subcommittee. They were unable to attend the Genetic Stocks Collection presentations on May 1 and 2, 1989 and thus could not participate in writing the report.)

APPENDIX F

LIST OF GENETIC STOCK COLLECTIONS FOR CROP SPECIES

Barley

Dr. T. Tsuchiya
College of Agricultural Sciences
Department of Agronomy
Colorado State University
Ft. Collins, Colorado 80523
(303) 491-6501

Cotton

Dr. Russell J. Kohel
USDA, ARS, Cotton and Grain Crops
Genetics Research
P.O. Drawer DN
College Station, Texas 77841
(713) 260-9311

Crucifer

Dr. Paul H. Williams
Department of Plant Pathology
University of Wisconsin
Madison, Wisconsin 53706
(608) 262-6496

Curcurbits

Dr. Richard W. Robinson
New York State Agricultural
Experiment Station
Hedrick Hall
Geneva, New York 14456
(315) 787-2237

Lettuce

Dr. E. J. Ryder
USDA/ARS
1636 East Alisal Street
Salinas, California 93905
(408) 443-2253

Maize

Dr. Earl Patterson
Department of Agronomy
Turner Hall
University of Illinois
Urbana, Illinois 61801
(217) 333-3420

Peanut

Dr. C. E. Simpson
Texas A&M University
P.O. Box 292
Stephenville, Texas 76401
(817) 968-4144

Peas

Dr. James R. McFerson
Regional Plant Introduction Station
New York State Agricultural
Experiment Station
Geneva, New York 14456-0462
(315) 787-2393

Phaseolus

Dr. Michael H. Dickson
Department of Seed and
Vegetable Sciences
New York State Agricultural
Experiment Station
Geneva, New York 14456
(315) 787-2222

Rice

Dr. R. H. Dilday
USDA, ARS
Rice Research & Extension Center
P. O. Box 287
Stuttgart, Arizona 72160
(501) 673-2661

Sorghum

Dr. Keith F. Schertz
USDA/ARS
Department of Soil and Crop Science
Texas A&M University
College Station, Texas 77843
(409) 260-9252

Soybean

Dr. Randall L. Nelson
USDA/ARS
University of Illinois
1102 S. Goodwin
Urbana, Illinois 61801
(217) 333-1117

Dr. Reid G. Palmer
USDA, ARS
G-301 Agronomy
Iowa State University
Ames, Iowa 50011
(515) 294-7378

Sugarbeet

Dr. Devon L. Doney
USDA/ARS
North Dakota State University
Walster Hall
Fargo, North Dakota 58105
(701) 237-8151

Sunflower

Dr. Gerald Seiler
USDA/ARS
Department of Agronomy
North Dakota State University
Fargo, North Dakota 58105
(701) 239-1380

Tomato

Dr. Charles M. Rick, Jr.
Department of Vegetable Crops
University of California
Davis, California 95616
(916) 752-1737

Wheat

Dr. J. P. Gustafsen
108 Curtis Hall
University of Missouri
Columbia, Missouri 65201
(314) 882-7318

THE GERMPLASM RESOURCES INFORMATION NETWORK - GRIN

A BRIEF HISTORY AND INTRODUCTION

UNITED STATES DEPARTMENT OF AGRICULTURE

AGRICULTURAL RESEARCH SERVICE

PLANT SCIENCE INSTITUTE

GERMPLASM RESOURCES INFORMATION NETWORK

March 1989

by

Database Management Unit

WHAT IS GRIN?

Plant genetic resources within the United States are managed by the National Plant Germplasm System (NPGS). One of the major structural components of this system is the Germplasm Resources Information Network (GRIN). GRIN is a centralized computer database used to facilitate management and operation of the NPGS and to enhance communication to scientists regarding the location and characteristics of material they may wish to obtain for research purposes. The major purpose of the GRIN is to serve as a central repository of information concerning both major and minor aspects of plant genetic resources within the NPGS and to provide ready accessibility of this information to all users of the system.

WHY HAVE GRIN?

Importance of Genetic Resources

Plant genetic resources (or germplasm) is the raw material required by plant breeders for the development of new, superior crop varieties that can ensure a stable, plentiful supply of high quality food, feed, and fiber. Most of the plants on which the United States depends were introduced from other countries and developed to suit the environment in which they would be grown. The list of economically important native plants is very short and includes sunflowers, cranberries, blueberries, strawberries, pecans, hops, range grasses, conifers, and hardwoods.

There are large gaps in the base of genetic diversity of some crops, particularly the wild species and primitive varieties. These rich sources of variation may contain genes for disease and insect resistance and other desirable traits, but in many areas of the world, these sources of diversity are rapidly being depleted, displaced, or abandoned. Once lost, these sources will never again be available to mankind. The need for this diversity becomes apparent when the genetic vulnerability of present American monoculture is measured against the constant battle against plant pathogens and pests.

Plant Introductions and the NPGS

The American Government recognized the need for a continuing search for more adaptable crops early in its history. In 1891, American overseas consuls were urged to send useful plants back to the United States. From this start, the essential elements of the present plant germplasm system gradually developed. This system evolved into the NPGS. The goal of the NPGS is to provide, on a continuing, long term basis, the plant genetic diversity needed by farmers and public and private plant scientists to improve productivity of crops and minimize their vulnerability to biological and environmental stresses.

Minimizing crop losses through control of major stresses is far more difficult and costly than increasing the genetic diversity among varieties of a given crop. Therefore, an NPGS objective is to broaden the genetic diversity of a crop throughout its production areas by having that production come from an array of varieties, all productive, but each different from the others in its range of tolerance to one or more potential stresses. Collection and introduction of new germplasm is the first step toward achievement of this goal.

The NPGS now maintains over 400,000 accessions (samples) of germplasm in the form of seed and vegetatively propagated stocks. These accessions are primarily landraces and unimproved material from foreign sources. New accessions are added to the NPGS at a rate of 7,000 to 15,000 per year. The need for the actual accession to enter the germplasm system is paralleled by the need for information about the accession to be available to users of the system. The immense size of this system creates a challenge for information management. Difficulty of obtaining information, lack of uniformity concerning this information, and its overall poor treatment prompted the NPGS to integrate an information management system as a major structural component. GRIN was designed and developed to act as this system.

BRIEF HISTORY OF GRIN

A feasibility study was conducted during 1976-77 which investigated and identified the need for information management systems that enhanced information availability with regard to collection, conservation, distribution, and utilization of plant genetic resources within the NPGS. This study drew the following conclusions about the existing information management system: "An information system exists within the plant genetic resources community of the United States but this system lacks the organization, communication techniques, trained personnel, and funding to satisfy the requirement of the NPGS community." From this study the USDA Agricultural Research Service (ARS) recognized the critical need for a nationally unified information system to serve the diverse needs of the NPGS. GRIN was the result of this realization. The developmental phase of this system was the Germplasm Resources Information Project (GRIP). This project was established under a 5-year cooperative agreement with the Laboratory for Information Science in Agriculture (LISA) to develop a computer based information system.

Analysis of the diverse needs of the germplasm community identified two groups of information users within the NPGS. The suppliers consist of those who acquire, maintain, and distribute genetic resources and data. It includes curators and staff of the National Seed Storage Laboratory (NSSL), Regional Plant Introduction Stations (RPIS), and other crop specific or clonal sites. A second, or demand group, consists of plant breeders, scientists, and other researchers from both the public and private sectors who use the genetic resources and data.

Further analysis identified specific needs of both groups. Small-scale prototypes were then constructed to meet the needs of each group on the supply side as well as to verify their objectives. From the evaluation of these prototypes, a user oriented approach was selected for design development.

In the evaluation and design of the information system, a centralized computer (Table 1) was selected to optimize operational speed for all users. A database concept was adopted to enhance information by reducing redundancy and by relating most of the pertinent information about a particular accession—from its native habitat to the most recent characteristic and evaluation results. This centralization also allows researchers access to a more extensive collection of samples from which to choose. This reduces the possibility of overlooking a potentially valuable sample. Maintenance of information is supported through updates that are quickly available to everyone. The most accurate and current information is thus accessible without time-consuming written notifications. Information updates are made by individuals and organizations recognized as experts in their particular area(s). For instance, plant taxonomists monitor and maintain taxonomic nomenclature; the Plant Introduction Office (PIO) in Beltsville, Maryland, maintains information concerning origin and particulars about introduction; the RPIS and other germplasm repositories maintain viable samples and serve as points of contact for sample availability and characteristic information; and the breeders, growers, and researchers provide detailed evaluation information.

To make the data understandable and consistent at a national level, Crop Advisory Committees (CAC's) were developed simultaneously with GRIP. These committees, composed of crop experts from public and private sectors of the NPGS, develop evaluation and characterization criteria as well as descriptor lists and standard methods of measurement and reporting.

The completed design phase brought the transformation of GRIP to GRIN. On July 1, 1983, GRIN was transferred to ARS, USDA. The management and final development of the network are controlled by the GRIN Database Management Unit (DBMU), Germplasm Services Laboratory, Plant Sciences Institute, located at Beltsville, Maryland. Implementation of this system (called GRIN1) was finalized in February 1984.

Soon after the implementation phase was finished, it was realized the system design was not sufficient to adequately include all parts of the NPGS and to accommodate the vast array of activities presented by this system. Major shortcomings included: omission of three major functional groups from the original design; inadequate germplasm ordering and inventory procedures; the use of too few data fields for the description of an accession; the lack of fast computer access to heavily used data; and a lack of fexibility and overall efficiency of the database. Another assessment of the needs was performed during 1985, when germplasm collection site personnel from 12 locations were brought to Beltsville to finalize the updated design. Expansion of the GRIN1 design and other programming was concluded in March 1987. Features of the

enhanced database design (GRIN2) compared to GRIN1 are: conversion of PIO, taxonomic support, and NSSL from their autonomous computer systems to GRIN; provision for germination rules and results; a simplification of the adding, modifying, and deleting of germplasm collection sites and associated inventory records; a more flexible characteristic and evaluation data structure; inclusion of germplasm ordering capabilities; a revision of and increase in the number of data fields; a doubling in size of the GRIN1 database; the inclusion of a revised friendly menu system for collection site users; a provision for collection site users to load their own data into the database; the inclusion of a number of predefined data reports; and an upgrade of security that increased efficiency of the database.

HOW DOES GRIN WORK?

GRIN has three important functions to fulfill. First, it serves as a central repository for valuable genetic resources information that is accessible by the entire scientific community. Second, it is a means for the CACs to begin standardization of crop descriptors and evaluation information. Third, it provides a mechanism for each of the RPIS and other sites to manage daily inventory.

Anyone who can justify a need for accessing the GRIN database can request access by writing the GRIN DBMU. Login identification and an access code are then assigned by the DBMU. Also, documentation and instructions for use of the system are supplied. Foreign scientists affiliated with any International Agricultural Research Center may also gain access through this procedure.

The database is designed to permit flexibility to the users in storing and retrieving information. A network design is one that allows multiple paths to the data but has linkages that connect all the data together. Table 2 contains a list of all data fields or elements that compose the GRIN2 database.

The public user system is designed for the inexperienced user and accommodates most hardware types that have telecommunication capabilities. Public users are presented with a series of menus which offer quick, easy access to major GRIN features and database searches. The menus assist unfamiliar users in the basic features of GRIN. Experienced users are also free to exercise GRIN features beyond the limits of the options presented through the use of menus.

The maintenance of data within GRIN is the responsibility of various functional groups within the NPGS. The PIO has responsibility for maintaining accurate passport data, geographic acquisition, and geographic origin information. USDA taxonomists maintain all plant taxonomic information. PIO and the taxonomists are able to modify information in their respective areas, however, any user possesses retrieval access. The germplasm collection sites are permitted to modify some designated passport information. The RPIS and other participating germplasm collection sites are reponsible for maintaining accurate inventory and characteristic data for their

respective collection. Public users are permitted to retrieve and examine all information from the database.

The DBMU acts as the caretaker of GRIN. This responsibility includes the maintenance of: all computer application software (programs); the database management system (DBMS); and a liaison with computer operations (the Prime minicomputer). The majority of volume data loading and the compilation and writing of GRIN documentation also lies with the DBMU. Database access and system security are also important system management tasks.

Contact the following office for any additional information:

Database Manager
USDA/ARS/PSI/GRIN/DBMU
Room 130, Building 001, BARC-West
Beltsville, Maryland 20705

Phone: 301-344-1666
FTS: 8-344-1666

Table 1. Description of the central GRIN computer

Computer location	Prime 9955 mod II National Agricultural Library Beltsville, Maryland, USA
Specifications and hardware	32 bit processor/data path 32 megabytes of memory 12.3 billion bytes of peripheral disk space situated on 16 disk drives 2 - 9 track 800/1600/6250 bpi tape drive 3 - synchronous communication ports 60 - asynchronous communication ports
Software	DBMS (Prime, Inc.) FORMS (Prime, Inc.) FORTRAN IV FORTRAN 77 BASIC (standard and compiled) PL1G INFO (Henco, Inc) PRIMOS (Prime, Inc., operating system)
Terminals and communications	Uninhibited communication for any terminal conforming to RS-232 protocol. Computer operates at full duplex, as default or half. Dialup access available for the following baud rates: 300 1200 2400 9600 19200 Telenet access available at 4800 and 9600 bps.

Table 2. Definitions for GRIN database ares and records

A. Area	Type of information
Accession	Passport or introduction
Accession-accessories	Sample acquisition, sample secondary accession identifiers
Cooperator	Names, addresses and institutes that cooperate with the NPGS
Descriptor	Definitions for variables used as descriptor state (codes)
Germination-rules	Seed germination conditions and other rules for seed testing
Inventory	Germplasm collection supply site inventory date
Observation	Single and multiple descriptor characterization and evaluation results
Order	Germplasm order requests data
Standards	Crop common names, geographic country and region names
Study	Environmental data for locations of characterization and evaluation trials, literature citation for these trials
Taxonomy	Taxonomic names

B. Record	
Accession	Principle accession identifiers which allow unique sample identification and detailed passport data
Accession-group	Names and groups of secondary identifiers
Alternate ID	Alternate identifier for the inventory samples
Code	Definitions of code values
Common-name	Crop common names
Composite-observation	Data from characterization and evaluation trials for multiple descriptors
Cooperator	Specific data about individuals and organizations that cooperate with the NPGS
Cooperator-group	Gives names to, and groups the cooperators
Descriptor	Detailed characterization and evaluation descriptor definitions
Descriptor-studied	Relates observation records to study and environment records
Distribution	Indicates a species presence in a geographical region
Environment	Environmental and location data for characterization and evaluation trials
Genus	Standard genus and family names

Germination-results	Results of sample germination tests
Germination-rule-authority	Relates species names with germination rules
Germination-rule	Guidelines for species germination tests
Geographic	Standard country and region names
Glossary	Keywords from publications concerning on characterization and evaluation trials
Inventory	Spatial locations, physical requirements for seed germination, etc.; basic data for daily germplasm management
Inventory-group	Group names for relating inventory records
Inventory-group-member	Relates inventory records with inventory groups
Keyword	Relates keywords to publications
Membership	Relates cooperators with cooperator groups by membership status
Observation	Observed values obtained during characterization trials
Order-items	Quantities and number of samples shipped per order
Order	Germplasm sample order names, number of items, etc.
Previous-names	Details of taxonomic name changes
Publication	Publication details for characterization and evaluation trials
Range	Groups accessions by taxonomy and includes donor names and acquisition country
Research-crop	Groups accessions by crop name for characterization purposes
Secondary-identifier	All accession identifiers except the primary identifier
Site	Collection site names and addresses
Site-crop	Groups inventory records for distribution, replenishment, and inventory control
Study	General details on characterization and evaluation trials
Supplier	Relates inventory samples with cooperators that supplied germplasm
Species-citation	Taxonomic literature citations
Species	Species and infraspecific taxonomic names
Synonym-citation	Taxonomic synonyms and reference citations
Taxonomic-literature	Complete names for taxonomic references
Taxonomic-synonym	Taxonomic synonym names